

¹ Fitness tracking reveals task-specific associations
² between memory, mental health, and exercise

³ Jeremy R. Manning^{1,*}, Gina M. Notaro^{1,2}, Esme Chen¹, and Paxton C. Fitzpatrick¹

⁴ ¹Dartmouth College, Hanover, NH

⁵ ²Lockheed Martin, Bethesda, MD

⁶ *Address correspondence to jeremy.r.manning@dartmouth.edu

⁷ October 4, 2021

⁸ **Abstract**

⁹ Physical exercise can benefit both physical and mental well-being. Different forms of exercise
¹⁰ (i.e., aerobic versus anaerobic; running versus walking versus swimming versus yoga; high-
¹¹ intensity interval training versus endurance workouts; etc.) impact physical fitness in different
¹² ways. For example, running may substantially impact leg and heart strength but only moderately
¹³ impact arm strength. We hypothesized that the mental benefits of exercise might be similarly
¹⁴ differentiated. We focused specifically on how different forms of exercise might relate to different
¹⁵ aspects of memory and mental health. To test our hypothesis, we collected nearly a century's
¹⁶ worth of fitness data (in aggregate). We then asked participants to fill out surveys asking them
¹⁷ to self-report on different aspects of their mental health. We also asked participants to engage in
¹⁸ a battery of memory tasks that tested their short and long term episodic, semantic, and spatial
¹⁹ memory. We found that participants with similar exercise habits and fitness profiles tended to
²⁰ also exhibit similar mental health and task performance profiles.

²¹ **Introduction**

²² Engaging in physical activity (exercise) can improve our physical fitness by increasing muscle
²³ strength (Crane et al., 2013; Knuttgen, 2007; Lindh, 1979; Rogers and Evans, 1993), increasing bone
²⁴ density (Bassey and Ramsdale, 1994; Chilibeck et al., 2012; Layne and Nelson, 1999), increasing
²⁵ cardiovascular performance (Maiorana et al., 2000; Pollock et al., 2000), increasing lung capac-
²⁶ ity (Lazovic-Popovic et al., 2016) (although see Roman et al., 2016), increasing endurance (Wilmore
²⁷ and Knuttgen, 2003), and more. Exercise can also improve mental health (Basso and Suzuki, 2017;
²⁸ Callaghan, 2004; Deslandes et al., 2009; Mikkelsen et al., 2017; Paluska and Schwenk, 2000; Raglin,
²⁹ 1990; Taylor et al., 1985) and cognitive performance (Basso and Suzuki, 2017; Brisswalter et al.,
³⁰ 2002; Chang et al., 2012; Ettnier et al., 2006).

³¹ The physical benefits of exercise can be explained by stress-responses of the affected body tis-
³² sues. For example, skeletal muscles that are taxed during exercise exhibit stress responses (Morton
³³ et al., 2009) that can in turn affect their growth or atrophy (Schiaffino et al., 2013). By comparison,
³⁴ the benefits of exercise on mental health are less direct. For example, one hypothesis is that ex-
³⁵ ercise leads to specific physiological changes, such as increased aminergic synaptic transmission
³⁶ and endorphin release, which in turn act on neurotransmitters in the brain (Paluska and Schwenk,
³⁷ 2000).

³⁸ Speculatively, if different exercise regimens lead to different neurophysiological responses, one
³⁹ might be able to map out a spectrum of signalling and transduction pathways that are impacted
⁴⁰ by a given type, duration, and intensity of exercise in each brain region. For example, prior work
⁴¹ has shown that exercise increases acetylcholine levels, starting in the vicinity of the exercised
⁴² muscles (Shoemaker et al., 1997). Acetylcholine is thought to play an important role in memory
⁴³ formation (Palacios-Filardo et al., 2021, e.g., by modulating specific synaptic inputs from entorhinal
⁴⁴ cortex to the hippocampus, albeit in rodents). Given the central role of these medial temporal
⁴⁵ lobe structures play in memory, changes in acetylcholine might lead to specific changes in memory
⁴⁶ formation and retrieval.

⁴⁷ In the present study, we hypothesize that (a) different exercise regimens will have different,

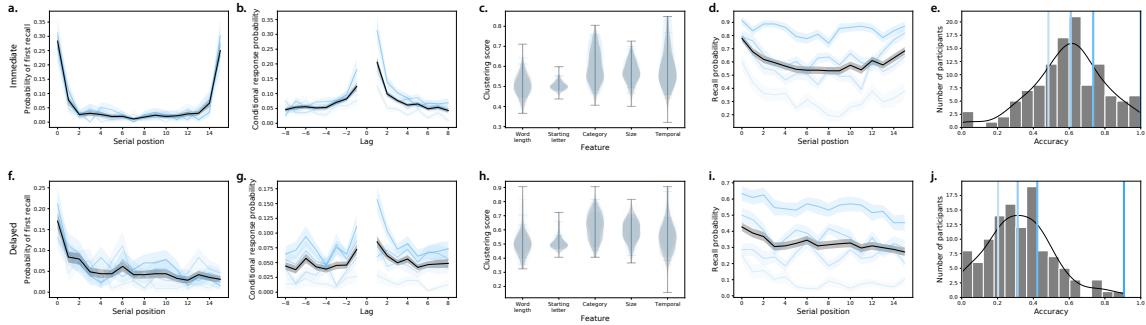


Figure 1: Free recall behavioral results.

48 quantifiable impacts on cognitive performance and mental health, and that (b) these impacts will
 49 be consistant across individuals. To this end, we collected a year of fitness tracking data from
 50 each of 113 participants. We then asked each participant to fill out a brief survey in which they
 51 self-evaluated several aspects of their mental health. Finally, we ran each participant through a
 52 battery of memory tasks, which we used to evaluate their memory performance along several
 53 dimensions. We examined the data for potential associations between memory, mental health, and
 54 exercise.

55 Results

56 Before testing our main hypothesis, we first examined the behavioral data from each memory
 57 task. We expected that the general trends and tendancies in the behavioral data would follow
 58 previously reported behaviors from similar tasks that had been utilized in prior work. We were also
 59 interested in characterizing the variability in task performance across participants. For example,
 60 if all participants exhibited near-identical behaviors or performance on a given task, we would be
 61 unable to identify how memory performance varied with mental health or exercise.

- 62 • characterizing behaviors (color by quartile and continue the color scheme in later figures-
 63 hue reflects task, shading reflects performance. white outline means immediate, black outline
 64 means delayed)

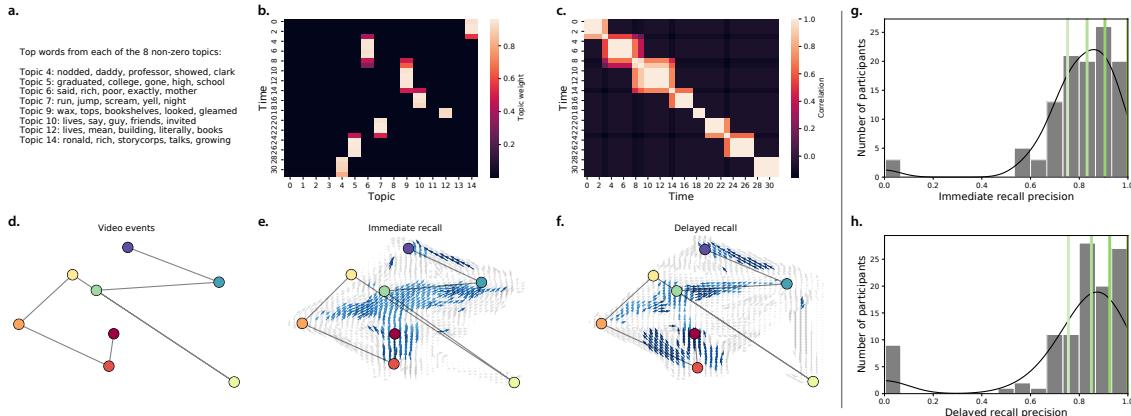


Figure 2: Naturalistic recall behavioral results.

- 65 – Free recall (immediate + delayed): pfr, lag-CRP, spc (color: recall performance). Figure 1.
- 66 – Naturalistic recall (immediate + delayed): reproduce a version of the sherlock movie/recall
67 trajectories (color: mean precision). Figure 2.
- 68 – Foreign language flashcards (immediate + delayed): p(correct) histogram (color: p(correct))
- 69 – Spatial learning: mean error by number of shapes (color: slope of line fit to errors as a
70 function of the number of shapes)
- 71 • Fitness info (break down by task performance, potentially separately for each task); also
72 separate out recent (raw) and recent versus baseline – color using same color scheme as
73 behavior figure
- 74 – activity (steps, zone minutes, floors/elevation)
- 75 – resting heart rate
- 76 – sleep
- 77 • exploratory analysis (correlations). Possibly make some sort of scatter matrix or pairplot-
78 rows/columns: tasks. Diagonal: histogram of performance metric for that task. Above-
79 diagonal entries: compare performance across tasks (by subject). Below-diagonal entries:
80 empty? density/2d histograms?

- 81 – Memory-memory
- 82 – fitness-fitness
- 83 – survey-survey
- 84 – (fitness + survey)-memory
- 85 • predictive analysis (regressions)
- 86 – Predict memory performance on held-out task from other tasks
- 87 – Predict memory performance on each task using fitness data
- 88 – Predict memory performance on each task using survey data
- 89 • Reverse correlations: look at recent changes versus baseline trends (color using same scheme
- 90 as behavior figures). Possibly
- 91 – Fitness profile that predicts performance on each task (barplots + timelines)
- 92 – Fitness profile for each survey demographic (barplots + timelines)
- 93 * Select out mental health demographics (based on meds, stress levels)

94 Discussion

- 95 • summarize key findings
- 96 • correlation versus causation
- 97 • what can vs. can't we know? we can identify correlations, but not causal direction– e.g. we
- 98 cannot know whether exercise *causes* mental changes versus whether people with particular
- 99 neural profiles might tend to engage in particular exercise behaviors. that being said, we *can*
- 100 separate out baseline tendencies (e.g., how people tend to exercise in general) versus recent
- 101 changes (e.g., how they happened to have exercised prior to the experiment).
- 102 • related work (exercise/memory, exercise/mental health), what this study adds

- 103 • future direction: towards customized physical exercise recommendation engine for optimiz-
104 ing mental health and mental fitness

105 **Methods**

106 We ran an online experiment using the Amazon Mechanical Turk platform. We collected data
107 about each participant’s fitness and exercise habits, a variety of self-reported measures concerning
108 their mental health, and about their performance on a battery of memory tasks. We mined the
109 dataset for potential associations between memory, mental health, and exercise.

110 **Experiment**

111 **Participants**

112 We recruited experimental participants by posting our experiment as a Human Intelligence Task
113 (HIT) on the Amazon Mechanical Turk platform. We limited participation to Mechanical Turk
114 Workers who had been assigned a “Masters” designation on the platform, given to workers who
115 score highly across several metrics on a large number of HITs, according to a proprietary algorithm
116 managed by Amazon. We further limited our participant pool to participants who self-reported that
117 they were fluent in English and regularly used a Fitbit fitness tracker device. A total of 160 workers
118 accepted our HIT in order to participate in our experiment. Of these, we excluded all participants
119 who failed to log into their Fitbit account (giving us access to their anonymized fitness tracking
120 data), encountered technical issues (e.g., by accessing the HIT using an incompatible browser,
121 device, or operating system), or who ended their participation prematurely, before completing the
122 full study. In all, 113 participants remained that contributed usable data to the study.

123 For their participation, workers received a base payment of \$5 per hour (computed in 15
124 minute increments, rounded up to the nearest 15 minutes), plus an additional performance-based
125 bonus of up to \$5. Our recruitment procedure and study protocol were approved by Dartmouth’s
126 Committee for the Protection of Human Subjects.

¹²⁷ **Gender, age, and race.** Of the 113 participants who contributed usable data, 77 reported their
¹²⁸ gender as female, 35 as male, and 1 chose not to report their gender. Participants ranged in age
¹²⁹ from 19–68 years old (25th percentile: 28.25 years; 50th percentile: 32 years; 75th percentile: 38
¹³⁰ years). Participants reported their race as White (90 participants), Black or African American (11
¹³¹ participants), Asian (7 participants), Other (4 participants), and American Indian or Alaska Native
¹³² (3 participants). One participant opted not to report their race.

¹³³ **Languages.** All participants reported that they were fluent in either 1 and 2 languages (25th
¹³⁴ percentile: 1; 50th percentile: 1; 75th percentile: 1), and that they were “familiar” with between 1
¹³⁵ and 11 languages (25th percentile: 1; 50th percentile: 2; 75th percentile: 3).

¹³⁶ **Reported medical conditions and medications.** Participants reported having and/or taking med-
¹³⁷ ications pertaining to the following medical conditions: anxiety or depression (4 participants),
¹³⁸ recent head injury (2 participants), high blood pressure (1 participant), bipolar (1 participant),
¹³⁹ hypothyroidism (1 participant), and other unspecified medications (1 participant). Participants
¹⁴⁰ reported their current and typical stress levels on a Likert scale as very relaxed (-2), a little relaxed
¹⁴¹ (-1), neutral (0), a little stressed (1), or very stressed (2). The “current” stress level reflected par-
¹⁴² ticipants’ stress at the time they participated in the experiment. Their responses ranged from -2
¹⁴³ to 2 (current stress: 25th percentile: -2; 50th percentile: -1; 75th percentile: 1; typical stress: 25th
¹⁴⁴ percentile: 0; 50th percentile: 1; 75th percentile: 1). Participants also reported their current level of
¹⁴⁵ alertness on a Likert scale as very sluggish (-2), a little sluggish (-1), neutral (0), a little alert (1),
¹⁴⁶ or very alert (2). Their responses ranged from -2 to 2 (25th percentile: 0; 50th percentile: 1; 75th
¹⁴⁷ percentile: 2). Nearly all (111 out of 113) participants reported that they had normal color vision,
¹⁴⁸ and 15 participants reported uncorrected visual impairments (including dyslexia and uncorrected
¹⁴⁹ near- or far-sightedness).

¹⁵⁰ **Residence and level of education.** Participants reported their residence as being located in the
¹⁵¹ suburbs (36 participants), a large city (30 participants), a small city (23 participants), rural (14 partic-
¹⁵² ipants), or a small town (10 participants). Participants reported their level of education as follows:

¹⁵³ College graduate (42 participants), Master's degree (23 participants), Some college (21 participants)
¹⁵⁴ High school graduate (9 participants), Associate's degree (8 participants), Other graduate
¹⁵⁵ or professional school (5 participants), Some graduate training (3 participants), or Doctorate (2
¹⁵⁶ participants).

¹⁵⁷ **Reported water and coffee intake.** Participants reported the number of cups of water and coffee
¹⁵⁸ they had consumed prior to accepting the HIT. Water consumption ranged from 0–6 cups (25th
¹⁵⁹ percentile: 1; 50th percentile: 3; 75th percentile: 4). Coffee consumption ranged from 0–4 cups (25th
¹⁶⁰ percentile: 0; 50th percentile: 1; 75th percentile: 2).

¹⁶¹ **Tasks**

¹⁶² Upon accepting the HIT posted on Mechanical Turk, the worker was directed to read and fill out
¹⁶³ a screening and consent form, and to share access to their anonymized Fitbit data via their Fitbit
¹⁶⁴ account. After consenting to participate and successfully sharing their Fitbit data, participants
¹⁶⁵ filled out a survey and then engaged in a series of memory tasks (Fig. 3). All stimuli and code for
¹⁶⁶ running the full Mechanical Turk experiment may be found [here](#).

¹⁶⁷ **Survey questions.** We collected the following demographic information from each participant:
¹⁶⁸ their birth year, gender, highest (academic) degree achieved, race, language fluency, and language
¹⁶⁹ familiarity. We also collected information about participants' health and wellness, including about
¹⁷⁰ their vision, alertness, stress, sleep, coffee and water consumption, location of their residence,
¹⁷¹ activity typically required for their job, and exercise habits.

¹⁷² **Free recall (Fig. 3a).** Participants studied a sequence of four word lists, each comprising 16 words.
¹⁷³ After studying each list, participants received an immediate memory test, whereby they were asked
¹⁷⁴ to type (one word at a time) any words they remembered from the just-studied list, in any order.

¹⁷⁵ Words were presented for 2 s each, in black text on a white background, followed by a 2 s blank
¹⁷⁶ (white) screen. After the final 2 s pause, participants were given 90 s to type in as many words
¹⁷⁷ as they could remember, in any order. The memory test was constructed such that the participant

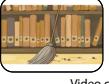
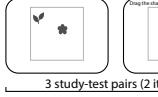
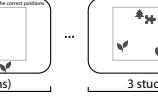
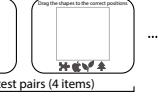
	Main task and immediate memory test				Delayed memory test
a.	1 Free recall 	Study words 	Memory test 		
b.	2 Naturalistic recall     Video clip plays	Memory tests  	Free response 		
c.	3 Foreign language flashcards 	Memory test 		7 	
d.	4 Spatial learning    3 study-test pairs (2 items) ... 3 study-test pairs (4 items) ... 3 study-test pair (7 items)			N/A	

Figure 3: Battery of memory tasks. **a. Free recall.** Participants study 16 words (presented one at a time), followed by an immediate memory test where they type each word they remember from the just-studied list. In the delayed memory test, participants type any words they remember studying, from any list. **b. Naturalistic recall.** Participants watch a brief video, followed by two immediate memory tests. The first test asks participants to write out what happened in the video. The second test has participants answer a series of multiple choice questions about the conceptual content of the video. In the delayed memory test, participants (again) write out what happened in the video. **c. Foreign language flashcards.** Participants study a sequence of 10 English-Gaelic word pairs, each presented with an illustration of the given word. During an immediate memory test, participants perform a multiple choice test where they select the Gaelic word that corresponds to the given photograph. During the delayed memory test, participants perform a second multiple choice test, where they select the Gaelic word that corresponds to each of a new set of photographs. **d. Spatial learning.** In each trial, participants study a set of randomly positioned shapes. Next, the shapes' positions are altered, and participants are asked to drag the shapes back to their previous positions. **All panels.** The gray numbers denote the order in which participants experienced each task or test.

178 could only see the text of the current word they were typing; when they pressed any non-letter
179 key, the current word was submitted and the text box they were typing in was cleared. This was
180 intended to prevent participants from retroactively editing their previous responses.

181 The word lists participants studied were drawn from the categorized lists reported in Ziman
182 et al. (2018). Each participant was assigned four unique randomly chosen lists (in a randomized
183 order), selected from a full set of 16 lists. Each chosen list was then randomly shuffled before
184 presenting the words to the participants.

185 Participants also performed a final delayed memory test where they were given 180 s to type
186 out any words they remembered from *any* of the 4 lists they had studied.

187 Recalled words within an edit distance of 2 (i.e., a Levenshtein Distance less than or equal to
188 2) of any word in the wordpool were “autocorrected” to their nearest match. We also manually
189 corrected clear typos or misspellings by hand (e.g., we corrected “hippoptumas” to “hippopota-
190 mus”, “zucinni” to “zucchini”, and so on). Finally, we lemmatized each submitted word to match
191 the plurality of the matching wordpool word (e.g., “bongo” was corrected to “bongos”, and so
192 on). After applying these corrections, any submitted words that matched words presented on the
193 just-studied list were tagged as “correct” recalls, and any non-matching words were discarded
194 as “errors.” Because participants were not allowed to edit the text they entered, we chose not to
195 analyze these putative “errors,” since we could not distinguish typos from true misrememberings.

196 **Naturalistic recall (Fig. 3b).** Participants watched a 2.5 minute video clip entitled “The Temple
197 of Knowledge.” The video comprises an animated story told to StoryCorps by Ronald Clark, who
198 was interviewed by his daughter, Jamilah Clark. The narrator (Ronald) discusses growing up
199 living in an apartment over Washington Heights branch of the New York Public Library, where his
200 father worked as a custodian during the 1940s.

201 After watching the video clip, participants were asked to type out anything they remembered
202 about what happened in the video. They typed their responses into a text box, one sentence at a
203 time. When the participant pressed the return key or typed any final punctuation mark (“.”, “!”, or
204 “?”) the text currently entered into the box was “submitted” and added to their transcript, and the

205 text box was cleared to prevent further editing of any already-submitted text. This was intended to
206 prevent participants from retroactively editing their previous responses. Participants were given
207 up to 10 minutes to enter their responses. After 4 minutes participants were given the option of
208 ending the response period early, e.g., if they felt they had finished entering all of the information
209 they remembered. Each participant's transcript was constructed from their submitted responses by
210 combining the sentences into a single document and removing extraneous whitespace characters.

211 Following this 4–10 minute free response period, participants were given a series of 10 multiple
212 choice questions about the conceptual content of the story. All participants received the same
213 questions, in the same order.

214 Participants also performed a final delayed memory test, where they carried out the free
215 response recall task a second time, near the end of the testing session. This resulted in a second
216 transcript, for each participant.

217 **Foreign language flashcards (Fig. 3c).** Participants studied a series of 10 English-Gaelic word
218 pairs in a randomized order. We selected the Gaelic language both for its relatively small number of
219 native speakers and for its dissimilarity to other commonly spoken languages amongst Mechanical
220 Turk Workers. We verified (via self report) that all of our participants were fluent in English and
221 that they were neither fluent nor familiar with Gaelic.

222 Each word's "flashcard" comprised a cartoon depicting the given word, the English word or
223 phrase in lowercase text (e.g., "the boy"), and the Gaelic word or phrase in uppercase text (e.g.,
224 "BUACHAIL"). Each flashcard was displayed for 4 s, followed by a 3 s interval (during which
225 the screen was cleared) prior to the next flashcard presentation.

226 After studying all 10 flashcards, participants were given a multiple choice memory test where
227 they were shown a series of novel photographs, each depicting one of the 10 words they had
228 learned. They were asked to select which (of 4 unique options) Gaelic word went with the given
229 picture. The 3 incorrect options were selected at random (with replacement across trials), and the
230 order in which the choices appeared to the participant were also randomized. Each of the 10 words
231 they had learned were tested exactly once.

232 Participants also performed a final delayed memory test, where they were given a second set of
233 10 questions (again, one per word they had studied). For this second set of questions participants
234 were prompted with a new set of novel photographs, and new randomly chosen incorrect choices
235 for each question. Each of the 10 original words they had learned were (again) tested exactly once
236 during this final memory test.

237 **Spatial learning (Fig. 3d).** Participants performed a series of study-test trials where they memo-
238 rized the onscreen spatial locations of two or more shapes. During the study phase of each trial,
239 a set of shapes appeared on the screen for 10 s, followed by 2 s of blank (white) screen. During the
240 test phase of each trial, the same shapes appeared onscreen again, but this time they were vertically
241 aligned and sorted horizontally in a random order. Participants were instructed to drag (using the
242 mouse) each shape to its studied position, and then to click a button to indicate that the placements
243 were complete.

244 In different study-test trials, participants learned the locations of different numbers of shapes
245 (always drawn from the same pool of 7 unique shapes, where each shape appeared at most one
246 time per trial). They first performed three trials where they learned the locations of 2 shapes; next
247 three trials where they learned the locations of 3 shapes; and so on until their last three trials, where
248 (during each trial) they learned the locations of 7 shapes. All told, each participant performed 18
249 study-test trials of this spatial learning task (3 trials for each of 2, 3, 4, 5, 6, and 7 shapes).

250 **Fitness tracking using Fitbit devices**

251 To gain access to our study, participants provided us with access to all data associated with their
252 Fitbit account from the year (365 calendar days) up to and including the day they accepted the HIT.
253 We filtered out all identifiable information (e.g., participant names, GPS coordinates, etc.) prior to
254 importing their data.

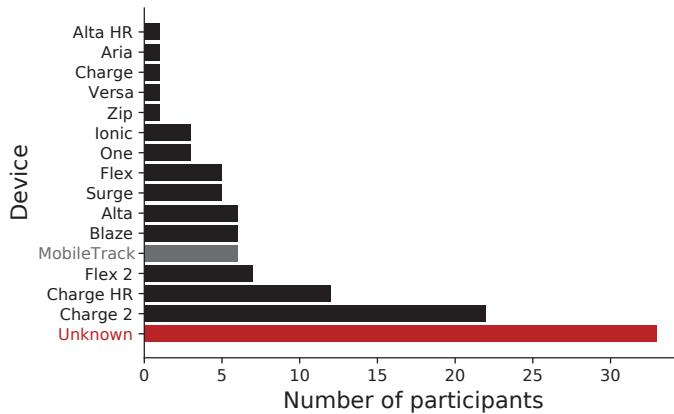


Figure 4: **Fitbit devices.** The bars indicate the numbers of participants whose fitness tracking data came from each model of Fitbit device. “MobileTrack” refers to participants who used smartphone accelerometer information to track their activity via the Fitbit smartphone app. “Unknown” denotes participants whose device information was not available from their available Fitbit data.

255 **Collecting and processing Fitbit data**

256 The fitness tracking data associated with participants’ Fitbit accounts varied in scope and duration
 257 according to which device the participant owned (Fig. 4), how often the participant wore (and/or
 258 synced) their tracking device, and how long they had owned their device. For example, while all
 259 participants’ devices supported basic activity metrics such as daily step counts, only a subset of
 260 the devices with heart rate monitoring capabilities provided information about workout intensity,
 261 resting heart rate, and other related measures.

262 Across all devices, we collected the following information: heart rate data, sleep tracking data,
 263 logged bodyweight measurements, logged nutrition measurements, Fitbit account and device
 264 settings, and activity metrics.

265 **Heart rate.** If available, we extracted all heart rate data collected by participants’ Fitbit device(s)
 266 and associated with their Fitbit profile. Depending on the specific device model(s) and settings, this
 267 included second-by-second, minute-by-minute, daily summary, weekly summary, and/or monthly
 268 summary heart rate information. These summaries include information about participants’ aver-
 269 age heart rates, and the amount of time they were estimated to have spent in different “heart rate

270 zones" (rest, out-of-range, fat burn, cardio, or peak, as defined by their Fitbit profile), as well as an
271 estimate of the number of estimated calories burned while in each heart rate zone.

272 **Sleep.** If available, we extracted all sleep data collected by participants' Fitbit device(s). Depend-
273 ing on the specific device model(s) and settings, this included nightly estimates of the duration
274 and quality of sleep, as well as the amount of time spent in each sleep stage (awake, REM, light, or
275 deep).

276 **Weight.** If available, we extracted any weight-related information affiliated with participants'
277 Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific device
278 model(s) and settings, this included their weight, body mass index, and/or body fat percentage.

279 **Nutrition.** If available, we extracted any nutrition-related information affiliated with participants'
280 Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific account
281 settings and usage behaviors, this included a log of the specific foods they had eaten (and logged)
282 over the past year, and the amount of water consumed each day.

283 **Account and device settings.** We extracted any settings associated with participants' Fitbit ac-
284 counts to determine (a) which device(s) and model(s) are associated with their Fitbit account, (b)
285 time(s) when their device(s) were last synced, and (c) battery level(s).

286 **Activity metrics.** If available, we extracted any activity-related information affiliated with par-
287 ticipants' Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific
288 device model(s) and settings, this included: daily step counts; daily amount of time spent in each
289 activity level (sedentary, lightly active, fairly active, or very active, as defined by their account
290 settings and preferences); daily number of floors climbed; daily elevation change; and daily total
291 distance traveled.

292 **Comparing recent versus baseline measurements.**

293 We were interested in separating out potential associations between *absolute* fitness metrics and
294 *relative* metrics. To this end, in addition to assessing potential raw (absolute) fitness metrics, we
295 also defined a simple measure of recent changes in those metrics, relative to a baseline:

$$\Delta_{R,B}m = \frac{B \sum_{i=1}^R m(i)}{R \sum_{i=R+1}^{R+B} m(i)},$$

296 where $m(i)$ is the value of metric m from $i - 1$ days prior to testing (e.g., $m(1)$ represents the value
297 of m on the day the participant accepted the HIT, and $m(10)$ represents the value of m 9 days prior
298 to accepting the HIT. Unless otherwise noted, we set $R = 7$ and $B = 30$. In other words, to estimate
299 recent changes in any metric m , we divided the average value of m taken over the prior week by
300 the average value of m taken over the 30 days before that.

301 **Exploratory correlation analyses**

302 We used a bootstrap procedure to identify reliable correlations between different memory-related,
303 fitness-related, and demographic-related variables. For each of $N = 1000$ iterations, we selected
304 (with replacement) a sample of 113 participants to include. This yielded, for each iteration, a
305 sampled “data matrix” with one row per sampled participant and one column for each measured
306 variable. When participants were sampled multiple times in a given iteration, as was often the case,
307 this matrix contained duplicate rows. We used a round-robin imputation procedure to estimate the
308 values of any missing features (Buck, 1960). Next, we computed the Pearson’s correlation between
309 each pair of columns. This yielded, for each pair of columns, a distribution of N bootstrapped
310 correlation coefficients. If fewer than 95% of the coefficients for a given pair of columns had the
311 same sign, we excluded the pair from further analysis and considered the expected correlation
312 between those columns to be undefined. If $\geq 95\%$ of the coefficients for a given pair of columns

313 had the same sign, we computed the expected correlation coefficient as:

$$\mathbb{E}_{i,j}[r] = \tanh\left(\frac{1}{N} \sum_{n=1}^N \tanh^{-1}(\text{corr}(m(i)_n, m(j)_n))\right),$$

314 where $m(x)_n$ represents column x of the bootstrapped data matrix for iteration n , \tanh is the
315 hyperbolic tangent, and \tanh^{-1} is the inverse hyperbolic tangent.

316 **Regression-based prediction analyses**

317 Following our exploratory correlation analyses, we used an analogous bootstrap procedure to iden-
318 tify subsets of memory-related, fitness-related, and demographic-related variables that predicted
319 (non-overlapping) subsets of other variables. For example, we tested whether a combination of
320 fitness-related variables could predict a combination of memory-related variables, and so on.

321 We used the same bootstrap procedure described above (used in our exploratory correlation
322 analyses) to generate $N = 1000$ bootstrapped data matrices whose rows reflected sampled partici-
323 pants and whose columns reflected different measured variables.

324 We grouped variables according to whether they were memory-related, fitness-related, or
325 demographic-related. For each bootstrap iteration, we divided the rows of that iterations data
326 matrix into training and test sets. The assignments of rows to these two sets was random, subject
327 to the constraint that any duplicated rows in the data matrix (i.e., reflecting a single participant who
328 had been sampled multiple times) was always assigned to either the training *or* the test set—i.e.,
329 duplicated rows could not appear in both the training and the test sets. The training sets always
330 comprised 75% of the data, and the tests sets comprised the remaining 25% of the data.

331 Next, we fit a series of ridge regression models to the training data. Specifically, for each pairing
332 of memory, fitness, and demographic variables, we fit a single ridge regression model treating the
333 first variable group as the input features and the second variable group as the target features. For
334 example, one regression model used memory variables to predict fitness variables, and another
335 regression model used fitness variables to predict demographic variables, and so on. In total we
336 fit six regression models to each training dataset. We then applied the fitted models to the held-

337 out test dataset and computed the root mean squared deviation (RMSD) between the predicted
338 and observed values in the target features of the test dataset. We also examined the regression
339 weights assigned to each input feature. This yielded, for each regression model (across N bootstrap
340 iterations) a distribution of RMSD values and a distribution of weights for each input variable.

341 We constructed a “null” distribution by using the same procedure as above, but where the
342 columns in the test datasets were randomly permuted with each iteration (thereby breaking any
343 meaningful predictive information between the training and test data). We assessed the statistical
344 significance (p -values) of the observed RMSD values by computing the proportions of null RMSD
345 values that were less than the observed value. We also assessed the significance of the observed
346 regression weights using t -tests to compare the means of the observed versus null distributions of
347 weights.

348 Reverse correlation analyses

349 We sought to characterize potential associations between the history of participants’ fitness-related
350 activities leading up to the time they participated in a memory task and their performance on
351 the given task. For each fitness-related variable, we constructed a timeseries matrix whose rows
352 corresponded to timepoints (sampled once per day) leading up to the day the participant accepted
353 the HIT for our study, and whose columns corresponded to different participants. These matrices
354 often contained missing entries, since different participants’ Fitbit devices tracked fitness-related
355 activities differently. For example, participants whose Fitbit devices lacked heart rate sensors
356 would have missing entries for any heart rate-related variables. Or, if a given participant neglected
357 to wear their fitness tracker on a particular day, the column corresponding to that participant
358 would have missing entries for that day.

359 In addition to this set of matrices storing timeseries data for each fitness-related variable, we also
360 constructed a memory performance matrix, M , whose rows corresponded to different memory-
361 related variables, and whose columns corresponded to different participants. For example, one
362 row of the memory performance matrix reflected the average proportion of words (across lists)
363 that each participant remembered during the immediate free recall test, and so on.

Given a fitness timeseries matrix, F , we computed the weighted average and weighted standard error of the mean of each row of F , where the weights were given by a particular memory-related variable (row of M). For example, if F contained participants' daily step counts, we could use any row of M to compute a weighted average across any participants who contributed step count data on each day. Choosing a row of M that corresponded to participants' performance on the naturalistic recall task would mean that participants who performed better on the naturalistic recall task would contribute more to the weighted average timeseries of daily step counts. Specifically, for each row, t , of F , we computed the weighted average (across the S participants) as:

$$\bar{f}(t) = \sum_{s=1}^S \hat{m}(s)F(t,s),$$

where \hat{m} denotes the normalized min-max scaling of m (the row of M corresponding to the chosen memory-related variable):

$$\hat{m} = \frac{m}{\sum_{s=1}^S \hat{m}(s)},$$

where

$$\hat{m} = \frac{m - \min(m)}{\max(m) - \min(m)}$$

We computed the weighted standard error of the mean as:

$$\text{SEM}_m(f(t)) = \frac{\left| \sum_{s=1}^S (F(t,s) - \bar{f}(t)) \right|}{\sqrt{S}}.$$

When a given row of F was missing data from one or more participants, those participants were excluded from the weighted average for the corresponding timepoint and the weights (across all remaining participants) were re-normalized to sum to 1. The above procedure yielded, for each memory variable, a timeseries of average (and standard error of the mean) fitness tracking values leading up to the day of the experiment.

³⁸¹ **Acknowledgements**

³⁸² We acknowledge useful discussions with David Bucci, Emily Glasser, Andrew Heusser, Abigail
³⁸³ Bartolome, Lorie Loeb, Lucy Owen, and Kirsten Ziman. Our work was supported in part by
³⁸⁴ the Dartmouth Young Minds and Brains initiative. The content is solely the responsibility of the
³⁸⁵ authors and does not necessarily represent the official views of our supporting organizations. This
³⁸⁶ paper is dedicated to the memory of David Bucci, who helped to inspire the theoretical foundations
³⁸⁷ of this work. Dave served as a mentor and colleague on the project prior to his passing.

³⁸⁸ **Data and code availability**

³⁸⁹ All analysis code and data used in the present manuscript may be found [here](#).

³⁹⁰ **Author contributions**

³⁹¹ Concept: J.R.M. Experiment implementation and data collection: G.M.N. Analyses: G.M.N., E.C.,
³⁹² P.C.F., and J.R.M. Writing: J.R.M.

³⁹³ **Competing interests**

³⁹⁴ The authors declare no competing interests.

³⁹⁵ **References**

- ³⁹⁶ Bassey, E. J. and Ramsdale, S. J. (1994). Increase in femoral bone density in young women following
³⁹⁷ high-impact exercise. *Osteoporosis International*, 4:72–75.
- ³⁹⁸ Basso, J. C. and Suzuki, W. A. (2017). The effects of acute exercise on mood, cognition, neurophys-
³⁹⁹ iology, and neurochemical pathways: a review. *Brain Plasticity*, 2(2):127–152.

- 400 Brisswalter, J., Collardeau, M., and René, A. (2002). Effects of acute physical exercise characteristics
401 on cognitive performance. *Sports Medicine*, 32:555–566.
- 402 Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use
403 with an electronic computer. *Journal of the Royal Statistical Society*, 22(2):302–306.
- 404 Callaghan, P. (2004). Exercise: a neglected intervention in mental health care? *Psychiatric and*
405 *Mental Health Nursing*, 11(4):476–483.
- 406 Chang, Y. K., Labban, J. D., Gapin, J. I., and Etnier, J. L. (2012). The effects of acute exercise on
407 cognitive performance: a meta-analysis. *Brain Research*, 1453:87–101.
- 408 Chilibek, P. D., Sale, D. G., and Webber, C. E. (2012). Exercise and bone mineral density. *Sports*
409 *Medicine*, 19:103–122.
- 410 Crane, J. D., MacNeil, L. G., and Tarnopolsky, M. A. (2013). Long-term aerobic exercise is associated
411 with greater muscle strength throughout the life span. *The Journals of Gerontology: Series A*,
412 68(6):631–638.
- 413 Deslandes, A., Moraes, H., Ferreira, C., Veiga, H., Silveira, H., Mouta, R., Pompeu, F. A. M. S.,
414 Coutinho, E. S. F., and Laks, J. (2009). Exercise and mental health: many reasons to move.
415 *Neuropsychobiology*, 59:191–198.
- 416 Etnier, J. L., Nowell, P. M., Landers, D. M., and Sibley, B. A. (2006). A meta-regression to examine the
417 relationship between aerobic fitness and cognitive performance. *Brain Research: Brain Research*
418 *Reviews*, 52(1):119–130.
- 419 Knutgen, H. G. (2007). Strength training and aerobic exercise: comparison and contrast. *Journal of*
420 *Strength and Conditioning Research*, 21(3):973–978.
- 421 Layne, J. E. and Nelson, M. E. (1999). The effects of progressive resistance training on bone density:
422 a review. *Medicine and Science in Sports and Exercise*, 31(1):25–30.

- 423 Lazovic-Popovic, B., Zlatkovic-Svenda, M., Durmic, T., Djelic, M., Saranovic, D., and Zugic, V.
424 (2016). Superior lung capacity in swimmers: some questions, more answers! *Revista Portuguesa*
425 *de Pneumologia*, 22(3):151–156.
- 426 Lindh, M. (1979). Increase of muscle strength from isometric quadriceps exercises at different knee
427 angles. *Scandinavian Journal of Rehabilitation Medicine*, 11(1):33–36.
- 428 Maiorana, A., O'Driscoll, G., Cheetham, C., Collis, J., Goodman, C., Rankin, S., Taylor, R., and
429 Green, D. (2000). Combined aerobic and resistance exercise training improves functional capacity
430 and strength in CHF. *Journal of Applied Physiology*, 88(1565–1570).
- 431 Mikkelsen, K., Stojanovska, L., Polenakovic, M., Bosevski, M., and Apostolopoulos, V. (2017).
432 Exercise and mental health. *Maturitas*, 106:48–56.
- 433 Morton, J. P., Kayani, A. C., McArdle, A., and Drust, B. (2009). The exercise-induced stress response
434 of skeletal muscle, with specific emphasis on humans. *Sports Medicine*, 39:643–662.
- 435 Palacios-Filardo, J., Udakis, M., Brown, G. A., Tehan, B. G., Congreve, M. S., Nathan, P. J., Brown, A.
436 J. H., and Mellor, J. R. (2021). Acetylcholine prioritises direct synaptic inputs from entorhinal cor-
437 tex to CA1 by differential modulation of feedforward inhibitory circuits. *Nature Communications*,
438 12(5475):doi.org/10.1038/s41467-021-25280-5.
- 439 Paluska, S. A. and Schwenk, T. L. (2000). Physical activity and mental health. *Sports Medicine*,
440 29(3):167–180.
- 441 Pollock, M. L., Franklin, B. A., Balady, G. J., Chaltman, B. L., Fleg, J. L., Fletcher, B., Limacher, M.,
442 na, I. L. P., Stein, R. A., Williams, M., and Bazzarre, T. (2000). Resistance exercise in individuals
443 with and without cardiovascular disease. *Circulation*, 101:828–833.
- 444 Raglin, J. S. (1990). Exercise and mental health. *Sports Medicine*, 9:323–329.
- 445 Rogers, M. A. and Evans, W. J. (1993). Changes in skeletal muscle with aging: effects of exercise
446 training. *Exercise and Sport Sciences Reviews*, 21:65–102.

- 447 Roman, M. A., Rossiter, H. B., and Casaburi, R. (2016). Exercise, ageing and the lung. *European*
448 *Respiratory Journal*, 48:1471–1486.
- 449 Schiaffino, S., Dyar, K. A., Ciciliot, S., Blaauw, B., and Sandri, M. (2013). Mechanisms regulating
450 skeletal muscle growth and atrophy. *The febs Journal*, 280(17):4294–4314.
- 451 Shoemaker, J. K., Halliwill, J. R., Hughson, R. L., and Joyner, M. J. (1997). Contributions of
452 acetylcholine and nitric oxide to forearm blood flow at exercise onset and recovery. *Vascular*
453 *Physiology*, 273(5):2388–2395.
- 454 Taylor, C. B., Sallis, J. F., and Needle, R. (1985). The relation of physical activity and exercise to
455 mental health. *Public Health Reports*, 100(2):195–202.
- 456 Wilmore, J. H. and Knuttgen, H. G. (2003). Aerobic exercise and endurance. *The Physician and*
457 *Sportsmedicine*, 31(5):45–51.
- 458 Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is automatic
459 speech-to-text transcription ready for use in psychological experiments? *Behavior Research*
460 *Methods*, 50:2597–2605.